International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 6, Issue 2, Apr 2016, 11-16

PRIVACY PRESERVING CLUSTERING BY ADDING DIFFERENT NOISE COMPONENTS GENERATED FROM PROGRESSIONS TO DIFFERENT CONFIDENTIAL ATTRIBUTES

T. SUDHA[1] & P. NAGENDRA KUMAR[2]

[1]*Professor, Department of Computer Science, Sri Padmavathi Mahila University, Tirupati, Andhra Pradesh, India*

[2]*Research Scholar, Department of Computer Science, Vikrama Simhapuri University, SPSR Nellore, Andhra Pradesh, India*

ABSTRACT

Privacy Preserving Data mining has been emerged as the one of the most prominent research areas in recent days. In this paper, we proposed a method for privacy preserving clustering of data. The proposed method considers a relational table and then identifies the confidential and non confidential attributes in the relational table so that non confidential attributes are removed and only the confidential attributes are retained in the table. The values of different confidential attributes are perturbed by adding different noise values generated from progressions. Then cluster analysis is performed on the original data and perturbed data using K-means algorithm with varying number of clusters. The results obtained show that the mean squared error obtained from the original data is same as the mean squared error obtained from the perturbed data but the order of values differ due to the random selection of cluster centers.

KEYWORDS: Privacy Preserving Clustering, Arithmetic Progression, Geometric Progression, Harmonic Progression

Received: Jan 25, 2016; **Accepted:** Feb 01, 2016; **Published:** Feb 25, 2016; **Paper Id.:** IJCSEITRAPR201602

Original Article

INTRODUCTION

Data mining refers to the practice of examining large pre-existing databases in order to generate new information. The different functionalities of data mining are concept description, association analysis, classification, clustering, outlier analysis and evolution analysis. Clustering is the process of grouping objects based on the similarity of objects. Clustering has many applications such as customer behavior analysis, targeted marketing, forensics and bioinformatics. Privacy is defined in terms of a person having control over the extent, timing and circumstances of sharing oneself with others. The goal of privacy preserving clustering is to protect the underlying attribute values of objects subjected to cluster analysis. In doing so, the privacy of individuals would be protected. Let D be a relational database and C a set of clusters generated from D. The goal is to transform D into D' so that a transformation T when applied to D must preserve the privacy of individual records, so that the released database D' conceals the values of confidential attributes and the similarity between objects in D' must be the same as that one in D.

Progression refers to a sequence of numbers in which each term is related to its predecessor by a uniform law. Progressions are of three types. They are Arithmetic progression, Geometric progression and Harmonic progression. An Arithmetic Progression (AP) is a sequence of numbers such that the difference of any two successive members of the sequence is a constant. A finite sequence of the form a, a+d, a+2d, a+3d…….a + (n-1) d

is called an Arithmetic Progression of n terms. The real number *'a'* is called the first term of the arithmetic progression and the real number'd' is called the common difference of the arithmetic progression. If a=1 and d=2 then the arithmetic progression is1, 3, 5, 7, 9, 11…2n-1.A Geometric progression (GP) is a sequence of numbers such that the quotient of any two successive members of the sequence is constant. A finite sequence of the form a, ar, ar^2, ar^3…….$ar^{n-1}$is called a Geometric Progression of n terms. The real number *'a'* is called the first term of the geometric progression and the real number *'r'* is called the common ratio of the geometric progression. If a=1 and r=2 then the geometric progression is1, 2, 4, 8…….2^n-1. A Harmonic progression (HP) is a progression formed by taking the reciprocals of an arithmetic progression. A Harmonic Progression is a sequence in which each term is the harmonic mean of the neighboring terms. A finite sequence of the form $a, \frac{a}{1+d}, \frac{a}{1+2d} \ldots\ldots\ldots\ldots \frac{a}{1+kd}$ is called a Harmonic Progression. If a=1 and d=2 then the harmonic progression is $1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7} \ldots\ldots \frac{1}{1+2k}$. Progressions form a part of Combinatorics. The branch of mathematics concerning the study of finite or countable discrete structures is called Combinatorics. Combinatorics has many applications in mathematical optimization, computer science ergodic theory and statistical physics.

PROPOSED METHOD

Any relational table in a database is considered. Non confidential attributes are removed from the table and only confidential attributes are retained. A particular progression from among the three progressions such as Arithmetic progression, Geometric progression and Harmonic progression is considered. With any choice of value for the initial term of the series and any choice of value for the common difference term(in case of AP and HP) and common ratio(in case of GP),generate a series of values in the progression. The number of values generated in the progression is equal to the number of confidential attributes in the table. Add first value of the progression to the first confidential attribute, second value of the progression to the second confidential attribute and so on. Then K-means algorithm of MATLAB is used for performing cluster analysis on both the original data as well as the perturbed data. The results obtained on the perturbed data is same as the results performed on the original data.

IMPLEMENTATION OF PROPOSED METHOD

Consider the following Employee table which consists of four attributes such as Employee Number, Employee Name, Age and Salary.

Table 1: Employee Table

Employee Number	Employee Name	Age	Salary(Rs.)
1	XXXX	40	43000
2	YYYY	45	48000
3	ZZZZ	38	28000
4	PPPP	36	26000
5	QQQQ	29	7000
6	RRRR	27	9000

In this table, 'Employee Number' and 'Employee Name' are treated as Non confidential attributes and the attributes 'Age' and 'salary' are treated as confidential. In the above table, non-confidential attributes are removed using anonymity model and the confidential attributes are retained in the following table.

Table 2: Employee Table with only Confidential Attributes

Age	Salary(Rs.)
40	43000
45	48000
38	28000
36	26000
29	7000
27	9000

Data Perturbation Using Noise Generated from Arithmetic Progression (AP)

Consider the following arithmetic progression1, 3, 5, 7, 9, 11, 13…… Here a=1 , d=2 where 'a' is the first term and 'd' is the common difference .First two values are selected from the progression since there are two confidential attributes in our table and the first value is added to the first confidential attribute and second value is added to the second confidential attribute. Then the perturbed data obtained is shown in the following table.

Table 3: Employee Table with Perturbed Data Obtained by Adding Noise Generated from Arithmetic Progression

Age	Salary
41	43003
46	48003
39	28003
37	26003
30	7003
28	9003

Cluster analysis is performed on both the original data and the perturbed data using K-means algorithm of MATLAB with varying number of clusters and the results obtained are shown in the table below.

Table 4: Mean Squared Error Obtained by Performing Clustering on Original Data and Perturbed Data Obtained Due to Noise Added from Arithmetic Progression

Number of Clusters	Cluster Number	Mean Squared Error Obtained from Original Data	Mean Squared Error Obtained from Perturbed Data
K=2	1	1.0e+008*0.1250	1.0e+008*0.1250
K=2	2	1.0e+008*3.6500	1.0e+008*3.6500
K=3	1	1.0e+007*0.2000	1.0e+007*0.2000
K=3	2	1.0e+007*0.2000	1.0e+007*1.2500
K=3	3	1.0e+007*1.2500	1.0e+007*0.2000
K=4	1	1.0e+007*1.2500	0
K=4	2	1.0e+007*0.2000	2000002
K=4	3	0	2000002
K=4	4	0	0
K=5	1	0	0
K=5	2	0	0
K=5	3	0	0
K=5	4	1.0e+007*1.2500	2000002
K=5	5	0	0
K=6	1	0	0
K=6	2	0	0
K=6	3	0	0
K=6	4	0	0

Table 4: Contd.,			
K=6	5	0	0
K=6	6	0	0

Data Perturbation Using Noise Generated from Geometric Progression (GP)

Consider the following geometric progression1, 2, 4, 8…… Here a=1, r=2 where ‘a’ is the first term and ‘r’ is the common ratio. First two values are selected from the progression since there are two confidential attributes in our table and the first value is added to the first confidential attribute and second value is added to the second confidential attribute. Then the perturbed data obtained is shown in the following table

Table 5: Employee Table with Perturbed Data Obtained by Adding Noise Generated from Geometric Progression

Age	Salary
41	43002
46	48002
39	28002
37	26002
30	7002
28	9002

Cluster analysis is performed on both the original data and the perturbed data using K-means algorithm of MATLAB with varying number of clusters and the results obtained are shown in the table below.

Table 6: Mean Squared Error Obtained by Performing Clustering on Original Data and Perturbed Data Obtained Due to Noise Added from Geometric Progression

Number of Clusters	Cluster Number	Mean Squared Error Obtained from Original Data	Mean Squared Error Obtained from Perturbed Data
K=2	1	1.0e+008*0.0200	1.0e+008*0.1250
K=2	2	1.0e+008*3.5675	1.0e+008*3.6500
K=3	1	1.0e+008*3.5675	1.0e+008*3.5675
K=3	2	0	0
K=3	3	0	0
K=4	1	1.0e+007*0.2000	0
K=4	2	0	1.0e+007*1.2500
K=4	3	0	0
K=4	4	1.0e+007*1.2500	1.0e+007*0.2000
K=5	1	0	1.0e+007*1.2500
K=5	2	0	0
K=5	3	2000002	0
K=5	4	0	0
K=5	5	0	0
K=6	1	0	0
K=6	2	0	0
K=6	3	0	0
K=6	4	0	0
K=6	5	0	0
K=6	6	0	0

Data Perturbation Using Noise Generated from Harmonic Progression (HP)

Consider the following harmonic progression1, 1/3, 1/5, 1/7…… Here a=1, d=2 where ‘a’ is the first term and‘d’ is the common difference. First two values are selected from the progression since there are two confidential attributes in our table and the first value is added to the first confidential attribute and second value is added to the second confidential attribute. Then the perturbed data obtained is shown in the following table.

Table 7: Employee Table with Perturbed Data Obtained by Adding Noise Generated from Harmonic Progression

Age	Salary
41	43000.3333
46	48000.3333
39	28000.3333
37	26000.3333
30	7000.3333
28	9000.3333

Cluster analysis is performed on both the original data and the perturbed data using K-means algorithm of MATLAB with varying number of clusters and the results obtained are shown in the table below.

Table 8: Mean Squared Error Obtained by Performing Clustering on Original Data and Perturbed Data Obtained Due to Noise Added from Harmonic Progression

Number of Clusters	Cluster Number	Mean Squared Error Obtained from Original Data	Mean Squared Error Obtained From Perturbed Data
K=2	1	1.0e+008*0.1250	1.0e+008*0.1250
K=2	2	1.0e+008*3.6500	1.0e+008*3.6500
K=3	1	1.0e+007*0.2000	1.0e+007*0.2000
K=3	2	1.0e+007*0.2000	1.0e+007*1.2500
K=3	3	1.0e+007*1.2500	1.0e+007*0.2000
K=4	1	0	1.0e+007*1.2500
K=4	2	2000002	0
K=4	3	2000002	1.0e+007*0.2000
K=4	4	0	0
K=5	1	0	0
K=5	2	0	2000002
K=5	3	1.0e+007*1.2500	0
K=5	4	0	0
K=5	5	0	0
K=6	1	0	0
K=6	2	0	0
K=6	3	0	0
K=6	4	0	0
K=6	5	0	0
K=6	6	0	0

CONCLUSIONS

In this paper we have presented a method for privacy preserving clustering using progressions. A sample relational table is considered and then the non confidential attributes are removed and confidential attributes are retained. Then the confidential data is perturbed by adding noise values generated from different progressions. Clustering is performed on both the original data and perturbed data using K-means algorithm of MATLAB with varying number of

clusters. From the results obtained, it is clear that mean squared error obtained on the perturbed data is same as the mean squared error obtained on the original data but the order of values differ due to random selection of cluster centers. Hence privacy preserving clustering is achieved. Instead of adding random noise values to different confidential attributes, our method adds the noise values generated from progressions to confidential attributes. It is very easy to memorize the noise values generated from progressions if we want to restore the original data from the perturbed data.

REFERENCES

1. *R. Agrawal and R. Srikant (May 2000). "Privacy Preserving Data Mining" in ACM SIGMOD, pages 439-450*
2. *P. Bunn and R. Ostrovsky (2007). "Secure Two-Party K-means Clustering" in ACM conference on Computer and Communications Security, pages 486-497*
3. *W. Du and M. Atallah (2001). "Privacy Preserving Cooperative Statistical Analysis" in 17^{th} ACSAC, pages 102-112*
4. *S. Guha, N. Mishra, R. Motwani and L. O'Callaghan (2000). "Clustering Data Streams" in IEEE FOCS, pages 359-366*
5. *S. Jha, L. Kruger and P. McDanial (2005). "Privacy Preserving Clustering" in ESORICS, pages 397-417*
6. *A.K. Jain, M.N. Murty and P.J. Flynn (1999). "Data Clustering: A review" ACM Comput. Surv. 31(3), 264-323*
7. *Y. Lindell and B. Pinkas (2002). "Privacy Preserving Data Mining", Journal of Cryptology, 15(3), 177-206*
8. *S. Oliveira and O. R. Zaiane (2003). "Privacy Preserving Clustering by Data Transformation" in Proceedings of the 18^{th} Brazilian Symposium on Databases, pages 304-318*
9. *J. Vaidya and C. Clifton (2003). "Privacy Preserving k-means Clustering over Vertically Partitioned Data" in 9^{th} KDD.ACM Press*
10. *J. Han and M. Kamber (2000). "Data mining: Concepts and Techniques", Morgan Kaufmann publishers Inc*
11. *J.A. Hartigan (1975). "Clustering Algorithms", John Wiley and Sons, Inc.*